IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Applicants:	Ludmilla Cherkasova et al.	§	Art Unit:	2155
		§		
Serial No.:	10/801,793	§		
		§	Examiner:	Michael Young Won
Filed:	March 16, 2004	§		
		§		
Title:	System And Method For Determining A Streaming Media Server Configuration For Supporting Expected Workload In Compliance With At Least One Service Parameter	§	Docket No.	200401021-1 (HPC.0519US)

Mail Stop **Appeal Brief-Patents**
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPEAL BRIEF

Date of Deposit: December 12, 2008

I hereby certify that this correspondence is being transmitted electronically to the U.S. Patent Office on the date indicated above.

Janice Munoz

TABLE OF CONTENTS

REAL PARTY IN INTEREST

The Hewlett-Packard Development Company, LP, a limited partnership established under the laws of the State of Texas and having a principal place of business at 20555 S.H. 249 Houston, TX 77070, U.S.A. (hereinafter "HPDC"). HPDC is a Texas limited partnership and is a wholly-owned affiliate of Hewlett-Packard Company, a Delaware Corporation, headquartered in Palo Alto, CA. The general or managing partner of HPDC is HPQ Holdings, LLC.

RELATED APPEALS AND INTERFERENCES

There are no related appeals and interferences.

STATUS OF CLAIMS

The application was originally filed with claims 1-46. Claims 2, 27, 33 and 41 have been cancelled. Claims 1, 3-26, 28-32, 34-40 and 42-46 have been finally rejected and are the subject of this appeal.

STATUS OF AMENDMENTS

All amendments have been entered.

SUMMARY OF CLAIMED SUBJECT MATTER

At this point, no issue has been raised that would suggest that the words in the claims have any meaning other than their ordinary meanings. Nothing in this section should be taken as an indication that any claim term has a meaning other than its ordinary meaning.

The method of independent claim 1 includes receiving, into a capacity planning system, workload information representing an expected workload of client accesses of streaming media files from a site (block 1202 in Fig. 12; Specification, para. no. [0126], ll. 25-26, p. 43); receiving, into the capacity planning system, at least one service parameter that defines a desired service characteristic to be provided by a media server configuration under the expected workload and defines a desired service characteristic to be provided by a media server configuration during periods of degraded service under the expected workload (block 1203 in Fig. 12; Specification, para. no. [0043], ll. 5-19, p. 12; and Specification, para. [0126], ll. 25-26, p. 43); and determining, by the capacity planning system, for at least one server configuration, how many servers of the configuration(s) are to be included at the site for supporting the expected workload in compliance with the service parameter (block 1204 in Fig. 12; Specification, para. no. [0126], ll. 2-5, p. 44).

The method of claim 11 includes receiving, into a capacity planning tool, information about a first server configuration (block 1201 of Fig. 12; Specification, para. no. [0126], ll. 20-21, p. 43); receiving, into the capacity planning tool, workload information representing an expected workload of client accesses of streaming media files from a site (block 1202 in Fig. 12; Specification, para. no. [0126], ll. 25-26, p. 43); and receiving, into the capacity planning system, at least one performability parameter that defines a desired service characteristic to be provided by a media server configuration during non-compliant periods of operation under the expected workload (block 1203 in Fig. 12; Specification, para. no. [0043], ll. 5-19, p. 12; and Specification, para. [0126], ll. 25-26, p. 43). The capacity planning tool determines how many servers of the first server configuration(s) are to be included at the site for supporting the expected workload in compliance with the performability parameter (block 1204 in Fig. 12; Specification, para. no. [0126], ll. 2-5, p. 44).

The method of claim 22 includes receiving, into a capacity planning tool, workload information representing an expected workload of client accesses of streaming media files over a

period of time T (Specification, para. no. [0107], ll. 16-20, p. 36). The capacity planning tool determines, for at least one media server configuration under evaluation, an amount of overload encountered by the media server configuration(s) during each of a plurality of time intervals of the expected workload (Specification, para. no. [0107], ll. 9-10, p. 36). The capacity planning tool receives at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload (Specification, para. no. [0107], ll. 10-13, p. 36).

The method of claim 30 includes receiving, into a capacity planning tool, workload information identifying an expected workload of client accesses of streaming media files from a server over a period of time T (Specification, para. no. [0107], ll. 16-20, p. 36); and determining, by the capacity planning tool, an interval overload profile for a media server configuration under evaluation, wherein the interval overload profile specifies an amount of overload of the media server configuration for each of a plurality of time intervals of duration I of the expected workload, where I<T (Specification, para. no. [0107], ll. 9-10, p. 36). The capacity planning tool determines based at least in part on the interval overload profile, whether the media server configuration under evaluation supports the expected workload in compliance with defined service parameters that define service characteristics desired by a service provider, and the defined service parameters include at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload (Specification, para. no. [0107], ll. 9-20, p. 36).

The system of claim 36 includes means for receiving workload information representing an expected workload of client accesses of streaming media files from a site over a period of time T (a media profiler 202 of Fig. 2; Specification, para. no. [0052], ll. 15-26, p. 16); means for determining, for at least one media server configuration under evaluation, an amount of overload encountered by the at least one media server configuration during servicing each of a plurality of time intervals of the expected workload (a capacity planner 101 of Fig. 2; Specification, para. no. [0107], ll. 17-19, p. 16); and a means for receiving at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by at least one media server configuration (Specification, para. no. [0107], ll. 17-19, p. 16).

The system of claim 44 includes a media profiler operable to receive workload information for a service provider's site and generate a workload profile for a server

configuration under consideration for supporting the service provider's site (a media profiler 202 of Fig. 2; Specification, para. no. [0052], ll. 15-17, p. 16); and a capacity planner operable to receive the generated workload profile for the server configuration under consideration and determine how many servers of the server configuration are needed to provide a media server solution having sufficient capacity for supporting the site's workload in compliance with defined performability parameters that specify a desired limit on degradation of quality of service provided by the media server solution during periods of degraded service (a capacity planner 101 of Fig. 2; Specification, para. [0126], ll. 25-26, p. 43).

GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

A. **Whether Claims 1, 3-26, 28-32, 34-40 and 42-46 Are Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

1. **Whether Claim 1 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

2. **Whether Claim 3 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

3. **Whether Claim 11 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

4. **Whether Claim 22 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

5. **Whether Claim 30 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

6. **Whether Claim 36 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

7. **Whether Claim 44 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?**

ARGUMENT

A. Whether Claims 1, 3-26, 28-32, 34-40 and 42-46 Are Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

1. Whether Claim 1 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The method of independent claim 1 includes receiving, into a capacity planning system, workload information representing an expected workload of client accesses of streaming media files from a site; receiving, into the capacity planning system, at least one service parameter that defines a desired service characteristic to be provided by a media server configuration under the expected workload and defines a desired service characteristic to be provided by a media server configuration during periods of degraded service under the expected workload; and determining, by the capacity planning system, for at least one server configuration, how many servers of the server configurations are to be included at the site for supporting the expected workload in compliance with the service parameter.

Claim 1 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over U.S. Patent Application Publication No. 2004/0111509 A1 (hereinafter called 'Eilam") in view of U.S. Patent No. 5,890,162 (hereinafter called "Huckins"). In general, Eilam describes techniques for the dynamic allocation of servers to purportedly maximize the revenue of a server farm. Eilam discloses that Internet traffic follows seasonal patterns and forecasts a current workload in view of this configuration. The servers are proactively allocated to customers according to the expected workload, and the allocation involves the use of a Multi-Layer infrastructure service level agreement (MLISLA), which among other metrics, specifies a minimum number of servers and a maximum number of servers for a customer. *See, for example,* para. nos. 20-22 of Eilam. Huckins generally discloses a system and technique for streaming semantics from a server computer.

To make a determination under 35 U.S.C. § 103, several basic factual inquiries must be performed, including determining the scope and content of the prior art, and ascertaining the

differences between the prior art and the claims at issue. *Graham v. John Deere Co.*, 383 U.S. 1, 17, 148 U.S.P.Q. 459 (1965). Moreover, as the U.S. Supreme Court held, it is important to identify a reason that would have prompted a person of ordinary skill in the art to combine reference teachings in the manner that the claimed invention does. *KSR International Co. v. Teleflex, Inc.*, 127 S. Ct. 1727, 1741, 82 U.S.P.Q.2d 1385 (2007).

Independent claim 1 recites receiving, into a capacity planning system, at least one service parameter that defines a desired service characteristic to be provided by a media server configuration under an expected workload and also defines a desired service characteristic to be provided by a media server configuration during periods of degraded service under the expected workload. The Final Office Action contends that Eilam discloses these limitations and relies on paragraph no. [0021] of Eilam for this finding. Final Office Action, pp. 3 and 19.

Applicant respectfully submits that the above-described finding is in error, as the cited language fails to disclose at least a parameter that defines a desired service characteristic to be provided by a media server configuration during periods of degraded service under an expected workload. Paragraph no. [0021] of Eilam is reproduced below:

> The invention offers the following advantages over prior art DRSF systems: Customers' application performance is improved because servers are proactively allocated to customers according to the expected workload, instead of waiting for threshold events triggered by workload increase and resulting performance deterioration, The revenue of the farm is maximized due to the optimization algorithm consolidating the workload forecast and the contract information. The system is more stable, due to the forecasting methods that can differentiate between temporary fluctuations in workload and real workload changes. Finally, the invention supports a flexible contract structure that might be suitable to a wider variety of customers.

As can be seen, paragraph no. 21 of Eilam fails to contain any disclosure related to a desired service characteristic during a period of degraded service. The disclosure of a mere allocation of servers does not meet the expressly recited claim limitations. It is noted that although paragraph no. 25 of Eilam describes a guarantee level, which may "be an upper bound on the percentage of time units in which a server that is needed ... is not available for the customer." However, this guarantee level likewise fails to disclose a parameter that defines a desired service characteristic during a period of degraded surface.

Huckins does not disclose the above-described claim limitations that are missing from Eilam and is apparently not being relied on by the Examiner for this purpose. Final Office, p. 3.

Therefore, Applicant submits that even if, for purposes of argument, Eilam and Huckins could be hypothetically combined, the hypothetical combination of the references would not have led to the claimed subject matter, as neither reference discloses nor suggests a parameter that defines a desired service characteristic during periods of degraded surface. The Final Office Action further fails to provide a plausible reason why it would have been obvious for one of skill in the art to derive the missing claim limitations in view of Eilam and Huckins. Therefore, for at least the foregoing reasons, Applicant respectfully submits that the § 103 rejection of independent claim 1 is in error and should be reversed.

2. Whether Claim 3 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The method of claim 3 depends from claim 1 and recites that the service parameter defines a limit on the amount of degradation of service encountered during the periods of degraded service.

Dependent claim 3 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Eilam in view of Huckins. Claim 3 is patentable for at least the reason that this claim depends from an allowable claim for the reasons that are set forth above. Claim 3 is patentable for at least the additional, independent reasons that are set forth below.

Applicant respectfully submits that the § 103 rejection of claim 3 is deficient for at least the reason that the Final Office Action errs in its factual finding. More specifically, the Final Office Action contends that paragraph no. 22 of Eilam discloses a performability parameter that specifies a limit on the amount of degradation of service encountered during periods of degraded service. Final Office Action, p. 11. Paragraph no. 22 of Eilam is reproduced below:

> The present invention supports a flexible and complex Multi-Layer Infrastructure Service Level Agreement (MLISLA) structure (LAYER.sub.1, LAYER.sub.2, LAYER.sub.3, . . . , LAYER.sub.k, MAP). The mapping function MAP maps monitored metrics (or an aggregation of metrics) to a required range of servers; namely, MAP(Mt)=(MIN(Mt), MAX(Mt)), meaning that at least MIN(Mt) number of servers are required to support the current load Mt, while any server beyond MAX(Mt) servers is superfluous for this purpose. Therefore, if the number of servers that are allocated when the load is Mt is less than X, the servers are considered overloaded. When load is between X and Y, the servers are considered well-balanced, and when the load is greater than Y, the servers are considered under-loaded.

As can be seen from the cited language, paragraph no. 22 of Eilam fails to disclose a service parameter that specifies a limit on the amount of degradation of service encountered during a period of degraded service. The Examiner fails to explain, other than citing paragraph no. 22, why the cited language renders claim 3 obvious. For at least the reason that the Final Office Action fails to make any showing why one of skill in the art in possession of Eilam and Huckins would have derive the limitations that are set forth in claim 3, Applicant respectfully submits that the § 103 rejection of claim 3 is in error and should be reversed.

3. Whether Claim 11 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The method of claim 11 includes receiving, into a capacity planning tool, information about a first server configuration; receiving, into the capacity planning tool, workload information representing an expected workload of client accesses of streaming media files from a site; and receiving, into the capacity planning system, at least one performability parameter that defines a desired service characteristic to be provided by a media server configuration during non-compliant periods of operation under the expected workload. The capacity planning tool determines how many servers of the first server configuration are to be included at the site for supporting the expected workload in compliance with the performability parameter.

Claim 11 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Eilam in view of Huckins. However, Applicant respectfully submits that the Final Office Action errs in the § 103 rejection of claim 11 for at least the following reasons.

The method of independent claim 11 recites receiving into a capacity planning system, at least one performability parameter that defines a service characteristic to be provided by a media server configuration during non-compliant periods of operation under an expected workload.

The Final Office Action states that paragraph no. 21 of Eilam purportedly discloses these limitations. However, Applicant respectfully submits that the language cited by the Office Action generally discusses mapping performance metrics to determine a number of servers but fails to disclose specifying at least one performability parameter that defines a service characteristic during a non-compliant period of operation. As discussed above, paragraph no. 25 of Eilam discloses a guarantee level. However, this guarantee level applies only to compliant

periods and does not address the performance during a non-compliant period. Huckins fails to disclose the missing limitations. Based on the misapplication of Eilam to the claimed subject matter, it is respectfully submitted that the obviousness rejection is defective.

Therefore, Applicant respectfully submits that the Final Office Action fails to show why one of skill in the art in possession of Eilam and Huckins would have derived the subject matter that is set forth in claim 11. As such, the § 103 rejection of claim 11 is in error and should be reversed.

4. Whether Claim 22 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The method of claim 22 includes receiving, into a capacity planning tool, workload information representing an expected workload of client accesses of streaming media files over a period of time T. The capacity planning tool determines, for at least one media server configuration under evaluation, an amount of overload encountered by the media server configuration(s) during each of a plurality of time intervals of the expected workload; and the capacity planning tool receives at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

Claim 22 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Eilam in view of Huckins. Applicant respectfully submits that the Final Office Action errs in the § 103 rejection of claim 22 for at least the following reasons.

Claim 22 recites that the capacity planning tool receives at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under an expected workload.

In the § 103 rejection of claim 22, the Final Office Action refers to paragraph no. 34 of Eilam. Final Office Action, p. 6. However, this cited language merely discloses server allocation by a resource manager 101 and fails to disclose any type of monitoring of the amount of continuous overload encountered by a media server configuration of at least one performability parameter that defines a desired limit on such a continuous overload. Thus, for at least these reasons, the § 103 rejection of claim 22 is defective, as Eilam does not disclose the above-identified subject matter, contrary to the position taken by the Final Office Action.

Huckins, which is relied on by the Final Office Action for purposes of disclosing streaming, does not discuss continuous overloads or a limit on a continuous overload.

Therefore, Applicant respectfully submits that one of skill in the art in possession of Eilam and Huckins would not have derived the claimed invention. As such, Applicant respectfully submits that the § 103 rejection of claim 22 is in error and should be reversed.

5. Whether Claim 30 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The method of claim 30 includes receiving, into a capacity planning tool, workload information identifying an expected workload of client accesses of streaming media files from a server over a period of time T; and determining, by the capacity planning tool, an interval overload profile for a media server configuration under evaluation. The interval overload profile specifies an amount of overload of the media server configuration for each of a plurality of time intervals of duration I of the expected workload, where I<T; and the capacity planning tool determines based at least in part on the interval overload profile, whether the media server configuration under evaluation supports the expected workload in compliance with defined service parameters that define service characteristics desired by a service provider. The defined service parameters include at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

Claim 30 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Eilam in view of Huckins.

Claim 30 recites that the defined service parameters include at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

As discussed above in connection with amended independent claim 22, the hypothetical combination of Eilam and Huckins would not have led to the claimed subject matter, as neither reference discloses a service parameter that includes at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration, and the Final Office Action fails to provide a plausible reason why it would have been obvious for one of skill in the art to derive the missing claim limitations.

Thus, for a t least the foregoing reasons, Applicant respectfully submits that the § 103 rejection of claim 30 is in error and should be reversed.

6. Whether Claim 36 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The system of claim 36 includes means for receiving workload information representing an expected workload of client accesses of streaming media files from a site over a period of time T; means for determining, for at least one media server configuration under evaluation, an amount of overload encountered by the media server configuration(s) during servicing each of a plurality of time intervals of the expected workload; and a means for receiving at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by at least one media server configuration.

Independent claim 36 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Eilam in view of Huckins.

Claim 36 includes a means for receiving at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by at least one media server configuration and overcomes the § 103 rejection for at least the same reasons as independent claims 22 and 30.

More specifically, for the purported disclosure of the claimed means for receiving the performance parameter(s), the Final Office Action refers to Eilam's discussion of server allocation by a resource manager 101. However, the cited language fails to disclose any type of monitoring of the amount of continuous overload encountered by a media server configuration or at least one performability parameter that defines a desired limit on such a continuous overload. As such, for at least these reasons, the § 103 rejection of claim 36 is in error and should be reversed.

7. Whether Claim 44 Is Rendered Obvious under 35 U.S.C. § 103(a) as Being Unpatentable over U.S. Patent Application Publication No. 2004/0111509 (Eilam) in View of U.S. Patent No. 5,890,162 (Huckins)?

The system of claim 44 includes a media profiler operable to receive workload information for a service provider's site and generate a workload profile for a server

configuration under consideration for supporting the service provider's site; and a capacity planner operable to receive the generated workload profile for the server configuration under consideration and determine how many servers of the server configuration are needed to provide a media server solution having sufficient capacity for supporting the site's workload in compliance with defined performability parameters that specify a desired limit on degradation of quality of service provided by the media server solution during periods of degraded service.

Independent claim 44 stands rejected under 35 U.S.C. § 103(a) as being unpatentable over Eilam in view of Huckins.

Applicant respectfully submits that the § 103 rejection of claim 44 is in error, as the Final Office Action errs in the factual finding to support this rejection. In this regard, paragraph nos. [21] and [22] of Eilam, which are relied on by the Final Office Action, fail to disclose, as discussed above, a performability parameter that specifies a desired limit on the degradation of quality of service during periods of degraded service. Although Eilam discloses in paragraph no. 25 a guarantee level on a percentage of time units in which a server that is needed is not available, Eilam fails to contain disclosure related to specifying a limit on the degradation of service during periods of degraded service.

The Final Office Action further refers to paragraph no. 36 of Eilam and contends that the discussion of a resource controlled algorithm in this paragraph discloses the claimed performability parameter. Final Office Action, 10. However, this finding is in error, as Eilam fails to disclose that the resource control algorithm in paragraph no. 36 specifies or receives a specification of a limit on performance during a degraded period of service. Huckins, which is relied on by the Final Office Action for purposes of its streaming discussion, also fails to disclose any such limit or parameter.

Therefore, Applicant respectfully submits that one of skill in the art in possession of Eilam and Huckins would not have derived the limitations of independent claim 44. As such, Applicant respectfully submits that the § 103 rejection of claim 44 is in error and should be reversed.

Applicant respectfully requests that each of the final rejections be reversed and that the claims subject to this Appeal be allowed to issue.

Respectfully submitted,

Date: December 12, 2008

Fred G. Pruner, Jr., Reg. No. 40,779
TROP, PRUNER & HU, P.C.
1616 S. Voss Road, Suite 750
Houston, Texas 77057
713/468-8880 [Phone]
713/468-8883 [Fax]

CLAIMS APPENDIX

The claims on appeal are:

1. A method comprising:

receiving, into a capacity planning system, workload information representing an expected workload of client accesses of streaming media files from a site;

receiving, into said capacity planning system, at least one service parameter that defines a desired service characteristic to be provided by a media server configuration under the expected workload and defines a desired service characteristic to be provided by a media server configuration during periods of degraded service under the expected workload; and

determining, by said capacity planning system, for at least one server configuration, how many servers of said at least one server configuration to be included at said site for supporting the expected workload in compliance with said at least one service parameter.

3. The method of claim 1 wherein said at least one service parameter specifies a limit on the amount of degradation of service encountered during said periods of degraded service.

4. The method of claim 1 wherein said at least one service parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

5. The method of claim 1 wherein said at least one service parameter comprises a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

6. The method of claim 1 wherein said at least one service parameter comprises at least one basic capacity parameter.

7. The method of claim 6 wherein said at least one basic capacity parameter comprises at least one selected from the group consisting of:

a statistical demand guarantee that specifies a desired limit on the percentage of time that a media server configuration is overloaded under the expected workload, and

a utilization constraint that specifies a desired limit on the percentage of time that a media server configuration is at or near its capacity under the expected workload.

8. The method of claim 6 wherein said at least one basic capacity parameter comprises a statistical demand guarantee that specifies a desired limit on the percentage of time that a media server configuration is overloaded under the expected workload, and a utilization constraint that specifies a desired limit on the percentage of time that a media server configuration is at or near its capacity under the expected workload.

9. The method of claim 6 wherein said at least one service parameter further comprises at least one performability parameter that defines a desired limit on the amount of degradation of service encountered during said percentage of time that a media server configuration is overloaded under the expected workload.

10. The method of claim 6 wherein said at least one service parameter further comprises at least one performability parameter that defines a desired limit on the amount of continuous overload encountered at any given time by a media server configuration under the expected workload.

11. A method comprising:

receiving, into a capacity planning tool, information about a first server configuration;

receiving, into said capacity planning tool, workload information representing an expected workload of client accesses of streaming media files from a site;

receiving, into said capacity planning system, at least one performability parameter that defines a desired service characteristic to be provided by a media server configuration during non-compliant periods of operation under the expected workload; and

said capacity planning tool determining how many servers of said first server configuration to be included at said site for supporting the expected workload in compliance with said at least one performability parameter.

12. The method of claim 11 wherein said non-compliant periods of operation comprise periods of degraded performance in servicing said expected workload.

13. The method of claim 12 wherein said degraded performance is performance in which said media server configuration is unable to satisfy real-time constraints of at least one stream being served.

14. The method of claim 12 wherein said degraded performance is performance in which said media server configuration is unable to serve at least one stream so as to avoid interruptions in the presentation of such stream.

15. The method of claim 12 wherein said degraded performance results from overload of said media server configuration.

16. The method of claim 11 wherein said non-compliant periods of operation comprise periods of at least one node failure of a clustered media server configuration.

17. The method of claim 11 further comprising:

receiving, into said capacity planning system, at least one basic capacity parameter that defines a desired service characteristic to be provided by a media server configuration during compliant periods of operation under the expected workload.

18. The method of claim 17 wherein said compliant periods of operation comprise periods in which said media server configuration is not overloaded under the expected workload.

19. The method of claim 17 further comprising:

said capacity planning tool performing basic capacity planning to determine how many servers of said first server configuration to be included at said site for supporting the expected workload in compliance with said at least one basic capacity parameter.

20. The method of claim 19 further comprising:

said capacity planning tool determining how many servers of said first server configuration to be included at said site for supporting the expected workload in compliance with said at least one basic capacity parameter and said at least one performability parameter.

21. The method of claim 11 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of said media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

22. A method comprising:

receiving, into a capacity planning tool, workload information representing an expected workload of client accesses of streaming media files over a period of time T;

said capacity planning tool determining, for at least one media server configuration under evaluation, an amount of overload encountered by said at least one media server configuration during each of a plurality of time intervals of said expected workload; and

said capacity planning tool receiving at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

23. The method of claim 22 where each of said plurality of time intervals have a size I where I<T.

24. The method of claim 22 wherein beginning points of each of said plurality of time intervals are separated by a Step amount.

25. The method of claim 24 wherein said Step<I.

26. The method of claim 24 wherein each of said intervals has a duration of 1 hour and said Step is 1 minute.

28. The method of claim 22 wherein said capacity planning tool evaluates said amount of overload encountered by said at least one media server configuration during each of said plurality of time intervals to determine whether said at least one media server configuration satisfies said at least one performability parameter.

29. The method of claim 22 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

30. A method comprising:

receiving, into a capacity planning tool, workload information identifying an expected workload of client accesses of streaming media files from a server over a period of time T;

determining, by said capacity planning tool, an interval overload profile for a media server configuration under evaluation, wherein said interval overload profile specifies an amount of overload of said media server configuration for each of a plurality of time intervals of duration I of said expected workload, where I<T; and

said capacity planning tool determining based at least in part on the interval overload profile whether said media server configuration under evaluation supports the expected workload in compliance with defined service parameters that define service characteristics desired by a service provider, wherein said defined service parameters include at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by a media server configuration under the expected workload.

31. The method of claim 30 wherein beginning points of each of said plurality of time intervals are separated by a Step amount.

32. The method of claim 31 wherein said Step<I.

34. The method of claim 30 wherein said capacity planning tool evaluates said interval overload profile for said media server configuration under evaluation to determine whether said media server configuration under evaluation satisfies said at least one performability parameter.

35. The method of claim 30 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of a media server configuration, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration have failed.

36. A system comprising:

means for receiving workload information representing an expected workload of client accesses of streaming media files from a site over a period of time T;

means for determining, for at least one media server configuration under evaluation, an amount of overload encountered by said at least one media server configuration during servicing each of a plurality of time intervals of said expected workload; and

a means for receiving at least one performability parameter that defines a desired limit on the amount of continuous overload encountered by at least one media server configuration.

37. The system of claim 36 further comprising:

means for receiving information specifying duration of each of said time intervals.

38. The system of claim 36 where each of said plurality of time intervals have a duration I where I<T.

39. The system of claim 36 wherein beginning points of each of said plurality of time intervals are separated by a Step amount.

40. The system of claim 39 wherein said Step is smaller in duration than a duration I of each of said intervals.

42. The system of claim 36 further comprising:

means for evaluating the determined amount of overload encountered by said at least one media server configuration under evaluation for each of said plurality of time intervals to determine whether said at least one media server configuration under evaluation satisfies said at least one performability parameter.

43. The system of claim 36 wherein said at least one performability parameter comprises at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of performance degradation under regular system operation of said at least one media server configuration under evaluation, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server configuration under evaluation have failed.

44. A system comprising:

a media profiler operable to receive workload information for a service provider's site and generate a workload profile for a server configuration under consideration for supporting the service provider's site; and

a capacity planner operable to receive the generated workload profile for the server configuration under consideration and determine how many servers of said server configuration are needed to provide a media server solution having sufficient capacity for supporting the site's workload in compliance with defined performability parameters that specify a desired limit on degradation of quality of service provided by said media server solution during periods of degraded service.

45. The system of claim 44 wherein said periods of degraded service comprise periods in which said media server configuration is unable to serve at least one stream so as to avoid interruptions in the presentation of such stream.

46. The system of claim 44 wherein said defined performability parameters comprise at least one selected from the group consisting of:

a regular-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods of degraded service under regular system operation of said media server solution, and

a node-failure-mode overload constraint that specifies a desired limit on the amount of degradation in service that is encountered during periods in which one or more nodes of a clustered media server solution have failed.

EVIDENCE APPENDIX

None.

RELATED PROCEEDINGS APPENDIX

None.